# Winter 2022 Data Science Intern Challenge

Please complete the following questions, and provide your thought process/work. You can attach your work in a text file, link, etc. on the application page. Please ensure answers are easily visible for reviewers!

**Question 1:** Given some sample data, write a program to answer the following: [click here to access the required data set](https://docs.google.com/spreadsheets/d/16i38oonuX1y1g7C_UAmiK9GkY7cS-64DfiDMNiR41LM/edit#gid=0)

On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

1. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

Answer: A major problem of using AOV is that the person who is buying a pair or two of shoes is clumped together with larger businesses ordering stock. Large orders can heavily skew the data, even if most orders are from a consumer. The 16th entry (line 17) is an example of this offset. In addition, there are just some orders that include suspiciously expensive shoes (orders involving Store 78) that also offset the data.

1. What metric would you report for this dataset?

I think a more suitable method might be to look for the average shoe value (ASV). Using this metric means we don’t need to worry about orders with larger quantities ordered. In addition, I would observe the potential outliers by calculating the interquartile range and using the third and first quartiles. After observing outliers using the inter quartile range, I can see that shoe store 42 and shoe store 78 are outliers. While store 78 has $25725 shoes and must go, store 42 has $352 shoes; some what more reasonable even though it is quite a bit more expensive than the average shoe. I decided to provide two calculations; one that just gets rid of the outliers using the IQR, and one that does the same but includes store 42.

1. What is its value?

Not including store 42, the Average Shoe Value is 150.40. Including store 42, we get the answer 152.48

**Question 2:** For this question you’ll need to use SQL. [Follow this link](https://www.w3schools.com/SQL/TRYSQL.ASP?FILENAME=TRYSQL_SELECT_ALL) to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.

1. How many orders were shipped by Speedy Express in total?

SELECT COUNT(\*) "Total Orders"

FROM Orders O, Shippers S

WHERE O.ShipperID = S.ShipperID AND S.ShipperName = 'Speedy Express';

Output:

|  |
| --- |
| **Total Orders** |
| 54 |

1. What is the last name of the employee with the most orders?

SELECT LastName

FROM Employees E JOIN Orders O USING(EmployeeID)

GROUP BY EmployeeID

HAVING COUNT(\*) = (SELECT MAX("Total Orders")

FROM (SELECT COUNT(\*) "Total Orders"

FROM Employees E JOIN Orders O USING(EmployeeID)

GROUP BY EmployeeID))

Output:

|  |
| --- |
| **LastName** |
| Peacock |

1. What product was ordered the most by customers in Germany?

SELECT ProductName

FROM Customers C

JOIN Orders O USING(CustomerID)

JOIN OrderDetails OD USING(OrderID)

JOIN Products P USING(ProductID)

WHERE Country = 'Germany'

GROUP BY ProductID

HAVING COUNT(\*) = (SELECT MAX("Total Orders") FROM(

SELECT COUNT(\*) "Total Orders"

FROM Customers C

JOIN Orders O USING(CustomerID)

JOIN OrderDetails OD USING(OrderID)

JOIN Products P USING(ProductID)

WHERE Country = 'Germany'

GROUP BY ProductID))

|  |
| --- |
| **ProductName** |
| Gorgonzola Telino |

Output: